Data mining the WhopperCropper database: sorghum in Central Queensland.

Brendan Power[1, 2] and Howard Cox[1]

[1] Department of Primary Industries and Fisheries, PO Box 102, Toowoomba, QLD 4350, Australia,
[2] Author for correspondence Brendan.Power@dpi.qld.gov.au

Abstract

WhopperCropper is a database of pre-run APSIM simulations. Originally developed for the Northern Grains Region it is currently being extended to all major cropping regions throughout Australia. It contains a factorial combination of agronomic inputs (e.g. crop variety and N application) with resources (e.g. soil water at sowing), and when completed it will contain approximately 5 million long-term simulations.

Large amounts of data however, does not necessarily equate to more knowledge. This paper highlights one of the many possible applications of *data mining* WhopperCropper. It describes a scenario from the Central Queensland Farming Systems project where information available to the decision maker at time of planting was used to test its capacity to identify optimal crop management strategies at planting. Cross validation techniques were used to ensure the independence of the predictions from the training data set.

This study showed that despite low ratio of explained to un-explained variability, considerable economic gain can be achieved by using WhopperCropper to tailor management options to prevailing planting conditions for soil with low water holding capacities.

Keywords

APSIM, WhopperCroppper, Decision Analysis

Introduction

WhopperCropper consists of a database of APSIM simulations and a graphical user interface with which the user analyses the outputs from pre-defined scenarios in terms of yields, gross margins and associated risks. However, due to the large number of potential combinations, the use of WhopperCropper can became tedious and time consuming which may contribute to the misuse of this decision/discussion support software.

In this paper we describe a holistic approach of searching and analysing the Whopper Cropper data base using the case of a sorghum crop at Emerald in Central Queensland.

Method

The WhopperCropper database

The factors and the levels of each factor analysed in this study are summarised in Table 1. The predetermined factors known at planting include: soil type (PAWC), plant available water at planting, planting date, and information from a seasonal climate forecast (i.e. SOI phase at planting, Stone and Auliciems, 1992). Manageable agronomic factors include: row configuration, planting densities, maturity types and applied nitrogen fertiliser.

It is also sometimes possible to manipulate the time of sowing. However, because the date is often determined by an incident rain event and soil water and SOI phase are uniquely defined at each planting opportunity, sow date is considered an “A” type factor (i.e. known at planting) for the purposes of this study.

A 115 year APSIM simulation for each factorial combination of the factors described in Table 1 (excluding SOI phase) was performed. This resulted in a matrix with more than 1 million rows.

Table 1. Definition of factor types and their levels from the WhopperCropper database

Factor type	Names	Levels
A. Known at planting	PAWC (mm)	80, 120, 150, 190
	Soil water at sowing (% of PAWC)	1/3, 2/3, full
	Sow date	15 Sep, 15 Oct, 15 Nov, 15 Dec, 15 Jan
	SOI phase	1 to 5
B. Agronomic manageable	Row configuration	solid, single skip, double skip
	Planting density (plants/m)	4, 8, 12
	Maturity type	quick, medium, slow
	Available N at sowing	25, 50, 100, 150, 200, 250

Gross Margin

To properly evaluate the effects of changing management options based on current agro-climatic conditions, a simple economic analysis was used. The following costs and prices are appropriate for Emerald at the present time (Chudleigh, personal communication). The on-farm price for sorghum was assumed to be $130/t. The fallow and planting cost were $77.49/ha, and the cost to apply elemental nitrogen was $1.09/kg, and harvest cost was $87.51/ha. Soil fertility was assumed to be low (25kg N/ha). These assumption resulted in a break even yield of 0.68t/ha. It was assumed that this was the minimum yield required for a crop to be harvested and hence incurred planting and harvesting costs. Seasons with a yield less than the break even yield, only incurred planting and fallow costs.

Decision analysis

This paper simulates decisions a farmer makes when growing sorghum in CQ and using WhopperCropper to select best agronomic management options for a forthcoming unknown season. This was done using leave-one-out cross validation, which involves splitting the data into independent training and testing data sets (years). That is, for each year from 1890 to 2005, management options were chosen from simulations conducted from all other years. This is the most efficient use of the data and is the best indicator of the predictive abilities (Hastie et al, 2001) of WhopperCropper.

Using information available at planting time (factor types A. in Table 1) such as soil moisture and a seasonal outlook, combinations of agronomic manageable factor type (B) were chosen using the following two strategies;

- a) a speculative approach where management options are chosen to maximise economic return and
- b) a conservative strategy where probabilities of financial loss were minimised. If two or more management strategies had the same probability of a financial loss, the strategy returning the greater gross margin was chosen.

Two benchmark strategies were included for comparison purposes. They are;

- c) a non-adaptive approach where the agronomic management used is typical for the region and soil conditions
- d) a hypothetical approach using perfect knowledge, where management options were chosen to maximise gross margins for the year of interest.

All strategies from a) to d) were applied for each unique combination of resource factors and years, to determine the best approach for all possible sowing conditions.

Analysis of variance

Significant factors in the decision making process were selected from an analysis of variance using the statistical computing environment R (R Development Core Team, 2006). However, due to the costs of nitrogen dominating the variability in gross margins, an ANOVA was performed on both yields and, on the probability of achieving at least the 'break-even' yield (i.e. 0.68t/ha) using a binomial Generalised Linear Model (GLM).

The results from the ANOVAs were used to determine which of the factors in Table 1, were used as a predictor for gross margin. In addition to this, the resultant sum of squares, were interpreted graphically, depicting the variability in yields and probability of financial loss due to each factor.

Results

All factors in Table 1, including most interaction terms, significantly influenced both average yields and probability of crop failure (results omitted). Hence all factors were considered in the decision analysis.

Figures 1 a) and b) display the partitioning of variability in mean yields (Figure 1a) and risk (Figure 1b) for all resource and agronomic factors listed in Table 1. They show that most of the variability (61% and 73% for yields and risk) was unexplained, labelled as residual in the plots, by the planting conditions and management options included in this analysis. Mean yields were slightly more predictable than crop risk as shown by less unexplained variability.

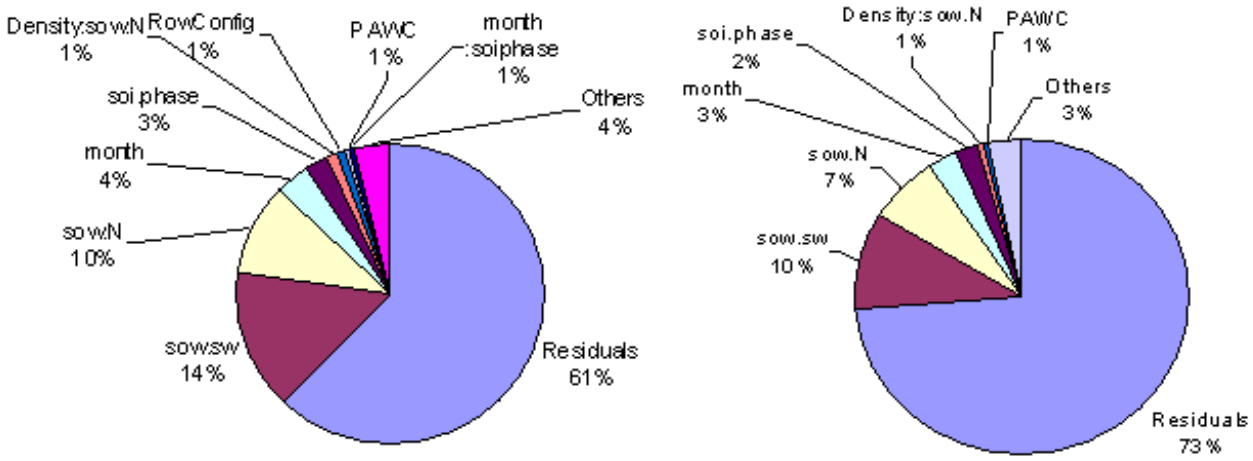
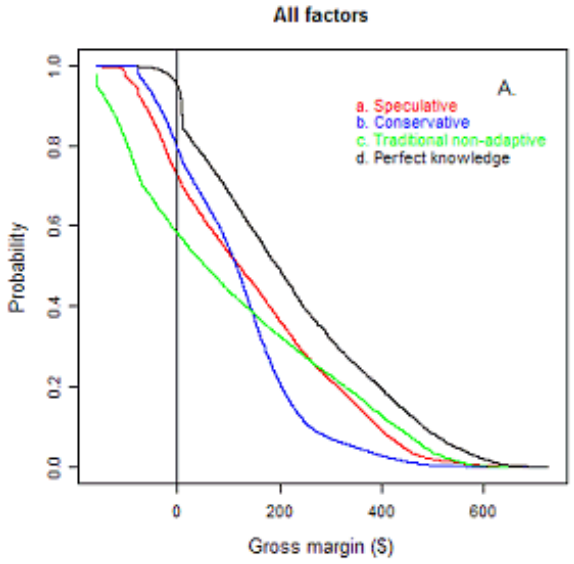


Figure 1. Partitioning of variability in mean yields a) and risk b).

Soil water at sowing accounted for most of the explained variability, followed by available nitrogen at sowing. This is to be expected from a water limited environment such as CQ. Surprisingly SOI phase at sowing accounted for more variability than planting density, maturity type and choice of row configuration. However, caution must be observed as these results are averaged over all factors and hence will vary from each particular scenario.

Figure 2A shows the probability of exceeding gross margin for all levels of each factor known at planting (A. type factors in Table 1). The benefits in utilising a data base such as WhopperCropper to select management strategies which optimise gross margins are evident. The adaptive strategies a) and b) (blue and red line) achieved an average increase in yield of $60/ha over the non-adaptive traditional management, strategy c) (green line), with considerably less risk.

These results however vary for each unique combination of starting conditions, as is apparent in the differences between plots B and C, in which soil PAWC was 120mm and 190mm respectively. The ability to predict the optimal management strategies at planting is abated by an increase in PAWC as shown by the curves for the adaptive strategies a) and b) being closer to perfect knowledge (black line) for shallow soils compared to deeper soils. This is likely due to the relative influence of in-crop precipitation. For the lower water holding soils, in-crop rain is more often lost to runoff or drainage and in greater quantities compared to soils with larger PAWCs. Therefore predicting an optimal management strategy at planting, when amount and timing of in-crop rain in unknown, is less successful when it has the most influence over yields.



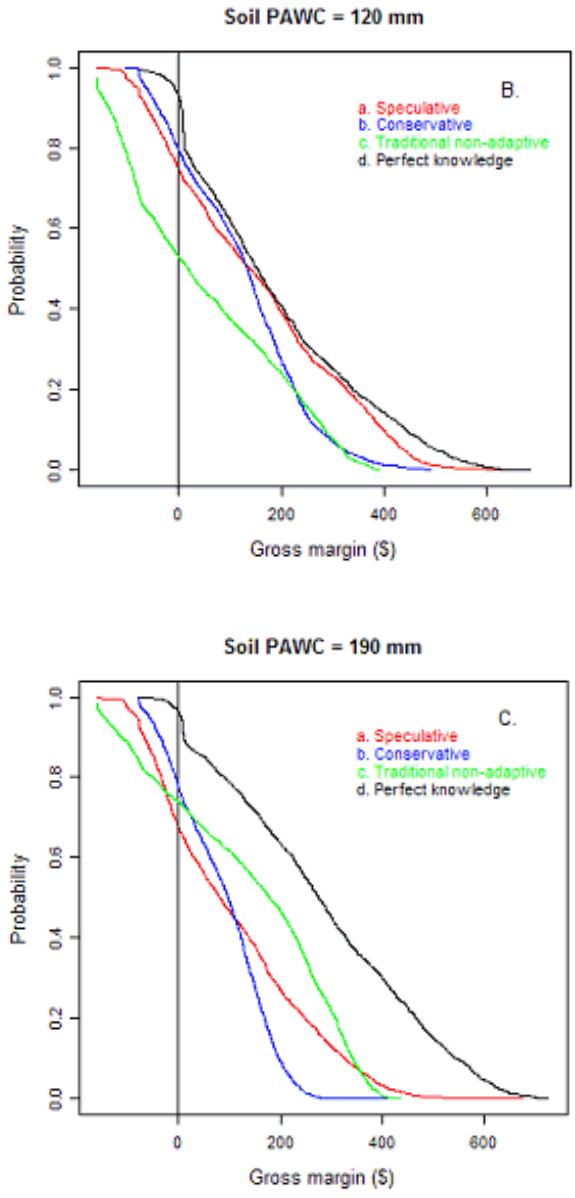


Figure 2. Cumulative probabilities showing management strategies a) speculative, b) conservative, c) traditional non-adaptive and d) perfect knowledge for all factors, soil PAWC 120mm and 190 mm.

Conclusion

Here we have shown that despite a low signal to noise ratio between sow conditions, agronomic management options and gross margins, there are benefits in using a tool such as WhopperCropper. However, here we tested the ability of historical simulations to predict the results of future independent simulations. However this does not necessarily account for the ability for simulations to predict reality.

The techniques used in this study to predict optimal management strategy are referred to as “Nearest Neighbour”. It is possible that other prediction techniques such as fitting a continuous model and using it to decide the best agronomic management may perform better because the fitting algorithm attempts to fit to signal and ignore noise, which doesn’t happen with “Nearest Neighbour” techniques. In addition

Nearest Neighbour fails to account for temporal correlation and non-stationary data as a result of climate change.

There is much future scope for this type of analysis, such as;

- Evaluating the usefulness of seasonal forecasts
- Multi-objective optimisation, e.g. minimising resource risk while maximising gross margin.
- Develop management strategies to deal with climate change
- Optimal strategies can be used to develop “rules-of-thumb”

Acknowledgments

This research was financially supported by QDPI and F, Land & Water Australia’s CVAP program and GRDC through the Central Queensland Farming Systems Project.

References

Chudleigh, F. Pers comm.

Hastie, Tibshirani and Friedman (2001), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M. and Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. *Europ. J. Agron*., **18**: 267-288.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Stone R. and Auliciems A.(1992). SOI phase relationships with rainfall in Eastern Australia, *International Journal of Climatology*, 12, 625-636.